Quadratic Semiparametric Von Mises Calculus

van der Vaart, A.W.; Tchetgen, E.; Robins, J.; Li, Lingling

published in
Metrika
2009

DOI (link to publisher)
10.1007/s00184-008-0214-3

document version
Publisher's PDF, also known as Version of record

Metrika (2009) 69:227–247
DOI 10.1007/s00184-008-0214-3

Quadratic semiparametric Von Mises calculus

James Robins · Lingling Li · Eric Tchetgen · Aad W. van der Vaart

Published online: 4 December 2008

Abstract We discuss a new method of estimation of parameters in semiparametric and nonparametric models. The method is based on U-statistics constructed from quadratic influence functions. The latter extend ordinary linear influence functions of the parameter of interest as defined in semiparametric theory, and represent second order derivatives of this parameter. For parameters for which the matching cannot be perfect the method leads to a bias-variance trade-off, and results in estimators that converge at a slower than $n^{-1/2}$-rate. In a number of examples the resulting rate can be shown to be optimal. We are particularly interested in estimating parameters in models with a nuisance parameter of high dimension or low regularity, where the parameter of interest cannot be estimated at $n^{-1/2}$-rate.

Keywords Von Mises calculus · Semiparametric models · Missing data · Tangent space · Influence function · Rate of convergence

1 Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution P_η with density p_η relative to a measure μ on a sample space $(\mathcal{X}, \mathcal{A})$, where the parameter η is known to belong to a subset H of a normed space. We wish to estimate the value $\chi(\eta)$ of a functional $\chi: H \to \mathbb{R}$ with the help of the observations $X_1, \ldots, X_n$. Our main

J. Robins · L. Li · E. Tchetgen
Department of Biostatistics and Epidemiology, School of Public Health, Harvard University, Cambridge, USA

A. W. van der Vaart (✉)
Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
e-mail: AW.van.der.Vaart@few.vu.nl

interest is in the situation of a semiparametric or nonparametric model, where H is an infinite-dimensional set, and the dependence $\eta \mapsto p_\eta$ is smooth.

This problem has been studied under the heading "semiparametric statistics" in the 1980s and 1990s. A theory of asymptotic lower bounds for "regular parameters" $\chi(\eta)$ based on Le Cam's concept of local asymptotic normality (Le Cam 1960) was developed starting with Koševnik and Levit (1976) and Pfanzagl (1982), and worked out for many examples in, among others, Begun et al. (1983), van der Vaart (1988) and Bickel et al. (1993). There are many examples of ad-hoc estimators that attain these bounds, and the behaviour of principled methods such as maximum likelihood (including its sieved and penalized variants) or estimating equations is understood to a certain extent (e.g., van der Vaart 1994; Murphy and van der Vaart 2000; Bolthausen et al. 2002; Wellner et al. 1993; van der Laan and Robins 2003).

Certain combinations of models $(P_\eta : \eta \in H)$ and parameter $\chi(\eta)$ possess structural properties that allow to estimate the parameter at $n^{-1/2}$-rate, no matter the size of the parameter set H. In this paper we are interested in the other situations, where the rate of estimation drops when the complexity of the model exceeds a certain limit. Such examples arise for instance when many covariates must be included in a model to correct for possible confounding in a causal study, or for modelling the probability that an individual is included in a sample in a study with missing observations. If simple (e.g., linear) models for the dependence on these covariates are not plausible, which is typical in epidemiological studies, then the resulting model must be taken so large that the usual methods fail. These methods typically focus on variance only, because the bias is negligible due to the structure of the model, or by explicitly assuming a "no-bias condition" (see Klaassen 1987; Murphy and van der Vaart 2000). In this paper we develop new methods that make a bias-variance trade-off when necessary.

These methods are based on quadratic estimating equations rather than the usual linear estimating equations.

Quadratic expansions for semiparametric models were previously investigated by Pfanzagl and Wefelmeyer (Pfanzagl 1985), but from the very different perspective of second order efficiency, i.e., the refinement of first order bounds by adding a lower order term. Our aim is to show that *second order influence functions* can be used for *first order inference*, because they permit balancing of bias and variance.

Following linear and quadratic is cubic, and so on. Extension of our approach to still higher orders is possible, but comes with many new complications. We shall pursue this elsewhere.

The paper is organized as follows. In Sect. 2 we review linear estimators from our current perspective. Next in Sect. 3 we introduce our new method of constructing quadratic estimators. This section has mostly a heuristic nature. In Sects. 4 and 5 we give rigorous constructions and results for two examples. The first is a classical theoretical example. The second is more extensive and concerns estimating a mean response when the response is not always observed.

Notation Let $\mathbb{P}_n$ and $\mathbb{U}_n$ denote the empirical measure and empirical U-statistic measure, viewed as an operators on functions: for given functions $f : \mathcal{X} \to \mathbb{R}$ and

$g: \mathcal{X}^2 \to \mathbb{R}$ these are given by

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \qquad \mathbb{U}_n g = \frac{1}{n(n-1)} \sum\sum_{1 \le i \ne j \le n} g(X_i, X_j).$$

We use the notation $\mathbb{U}_n f$ also for $f: \mathcal{X} \to \mathbb{R}$ a function of one argument, with the interpretation $\mathbb{U}_n f = \mathbb{P}_n f$. This is consistent with the given formulas if a function of one argument is considered as a function of 2 arguments that is constant in its second argument.

We write $P^n \mathbb{U}_n g = P^2 g$ for the expectation of $\mathbb{U}_n g$ if $X_1, \ldots, X_n$ are distributed according to the probability measure P. We also use the operator notation for the expectations of statistics in general.

We call a measurable function $g: \mathcal{X}^2 \to \mathbb{R}$ *degenerate* relative to P if $\int g(x_1, x_2)\, dP(x_i) = 0$ for $i = 1, 2$, and we call it *symmetric* if $g(x_1, x_2) = g(x_2, x_1)$ for every $x_1, x_2 \in \mathcal{X}$.

Given two functions $g, h: \mathcal{X} \to \mathbb{R}$ we write $g \times h$ for the function $(x, y) \mapsto g(x)h(y)$. Such tensor products functions are degenerate if both functions f and g have mean zero. The corresponding notation $P \times Q$ of two measures P and Q gives the product measure.

2 Linear estimator

Given an initial estimator $\hat{\eta}$ of η, the plug-in estimator $\chi(\hat{\eta})$ is typically a consistent estimator of the parameter of interest $\chi(\eta)$, but it may not be a good estimator. In particular, if $\hat{\eta}$ is a general purpose estimator, not specially constructed to yield a good plug-in, then $\chi(\hat{\eta})$ will often have a suboptimal precision. To gain insight in this situation we assume that the parameter permits a Taylor expansion of the form

$$\chi(\eta) = \chi(\hat{\eta}) + \chi'_{\hat{\eta}}(\eta - \hat{\eta}) + O\left(\|\eta - \hat{\eta}\|^2\right). \tag{1}$$

Such an expansion suggests that the plug-in estimator will have an error of the order $O_P\left(\|\eta - \hat{\eta}\|\right)$, unless the linear term in the expansion vanishes.

The expansion (1) also suggests that better estimators can be obtained by "estimating" the linear term in the expansion. To achieve this we assume a "generalized von-Mises representation" of the derivative of the form

$$\chi'_{\hat{\eta}}(\eta - \hat{\eta}) = \int \dot{\chi}_{\hat{\eta}}\, d(P_\eta - P_{\hat{\eta}}) = \int \dot{\chi}_{\hat{\eta}}\, dP_\eta + O\left(\|\eta - \hat{\eta}\|^2\right), \tag{2}$$

for some measurable function $\dot{\chi}_{\hat{\eta}}: \mathcal{X} \to \mathbb{R}$, referred to as an *influence function*. The second equality is valid if $\dot{\chi}_\eta$ is degenerate relative to P_η (i.e., $P_\eta \dot{\chi}_\eta = 0$) for every η, which can always be arranged by a recentering, as $\int 1\, d(P_\eta - P_{\hat{\eta}}) = 0$. The von-Mises representation and Eq. (1) suggest the "corrected plug-in estimator"

$$\chi(\hat{\eta}) + \mathbb{P}_n \dot{\chi}_{\hat{\eta}}. \tag{3}$$

This estimator should have an error of the order $O_P(n^{-1/2}) + O_P\left(\|\eta - \hat{\eta}\|^2\right)$, as the difference $(\mathbb{P}_n - P_\eta)\dot{\chi}_{\hat{\eta}}$ is "centered" and ought to have "variance" of the order $O(1/n)$.

We put "centered" and "variance" in quotes, because the randomness in the initial estimator $\hat{\eta}$ prevents a simple calculation of mean and variance. Empirical process theory can be used to show that the effect of replacing $\dot{\chi}_\eta$ by $\dot{\chi}_{\hat{\eta}}$ is negligible, if the class of functions $\dot{\chi}_\eta$ is not too rich. In the present paper we are interested in orders of magnitude only, and then a simpler approach is to split the sample and use separate observations to construct $\hat{\eta}$ and to construct $\mathbb{P}_n$. Then the orders can be justified by reasoning conditionally on the first sample, and it suffices that $\int \dot{\chi}_{\hat{\eta}}^2 \, dP_\eta$ remains bounded in probability.

Von Mises (1947) originally introduced the expansions that are named after him in order to investigate functionals of empirical distributions. The idea to use expansions (1) for estimation in nonparametric models occurs in Emery et al. (2000). Our situation is more involved, because we are interested in models $(P_\eta : \eta \in H)$ that are structured through a map $\eta \mapsto p_\eta$, and we are interested in a functional $\chi(\eta)$ of the parameter. In this situation a von-Mises type expansion can fail for two reasons. First a derivative χ'_η is by definition a continuous, linear map on the underlying normed space, and such maps may or may not be representable as an integral, depending on the normed space. Second, our von Mises expansion (2) represents this derivative as an integral relative to the distribution P_η and hence also involves the inverse map $P_\eta \to \eta$ from the distribution of the data to the parameter. We require representation through P_η, because this allows us to construct the estimator (3) by replacing P_η by the empirical distribution.

These issues are related to investigations in the theory of semiparametric models (see Koševnik and Levit 1976; Pfanzagl 1982; van der Vaart 1988; Bickel et al. 1993). These papers define a *tangent set* of a semiparametric model $(P_\eta : \eta \in H)$ as the set of functions $\dot{g}_\eta : \mathcal{X} \to \mathbb{R}$ obtainable as

$$\tfrac{1}{2}\dot{g}_\eta \sqrt{p_\eta} = \lim_{t \downarrow 0} \frac{\sqrt{p_{\eta_t}} - \sqrt{p_\eta}}{t},$$

where the limit is taken in the L_2-sense, and $t \mapsto \eta_t$ ranges over a collection of maps from $[0, 1] \subset \mathbb{R}$ to H for which the limit exists. Informally, a "tangent vector" $\dot{g}_\eta$ is just a *score function*

$$\dot{g}_\eta = \frac{\partial}{\partial t}_{|t=0} \log p_{\eta_t} = \frac{\frac{\partial}{\partial t}_{|t=0} p_{\eta_t}}{p_\eta}, \tag{4}$$

of a one-dimensional submodel $(P_{\eta_t} : t \geq 0)$ at $t = 0$, where $\eta_0 = \eta$. (Taking the derivative in the L_2-sense is appropriate for asymptotic information theory, but not necessarily so for the present heuristic discussion.) An *influence function* is defined as a measurable map $\dot{\chi}_\eta : \mathcal{X} \to \mathbb{R}$ such that, for all paths $t \mapsto \eta_t$ considered,

$$\frac{d}{dt}_{|t=0} \chi(\eta_t) = P_\eta \dot{\chi}_\eta \dot{g}_\eta. \tag{5}$$

It is not difficult to see that the latter influence function is the same as the influence function needed in the von-Mises expansion (2), if the various types of derivatives match up. (Note that the middle expression in (2) with η replaced by η_t and $\hat{\eta}$ by η expands to $P_\eta \dot{\chi}_\eta \dot{g}_\eta + o(t)$, as $p_{\eta_t} - p_\eta = t\dot{g}_\eta p_\eta + o(t)$.) Necessary and sufficient conditions for existence of an influence function in terms of the derivatives of the maps $\eta \mapsto \chi(\eta)$ and $\eta \mapsto p_\eta$ were investigated in van der Vaart (1991).

An influence function is not necessarily unique, as only its inner products with elements $\dot{g}_\eta$ of the tangent set matter. The projection of any influence function that is contained in the closed linear span of the tangent set is called the *efficient influence function* or *canonical gradient*, as it is the influence function of asymptotically efficient estimators. It minimizes the variance $\mathrm{var}_\eta \mathbb{P}_n \dot{\chi}_\eta$ over all influence functions.

3 Quadratic estimator

If the preliminary estimator $\hat{\eta}$ attains a rate of convergence $\|\hat{\eta} - \eta\| = o_P(n^{-1/4})$, then the plug-in estimator (3) attains a $n^{-1/2}$-rate of convergence. Typically this will require that the parameter set H is not too big. If the preliminary estimator is less precise, then the remainder term of the expansion (1) will dominate. This suggests to take the expansion further to

$$\chi(\eta) = \chi(\hat{\eta}) + \chi'_{\hat{\eta}}(\eta - \hat{\eta}) + \tfrac{1}{2}\chi''_{\hat{\eta}}(\eta - \hat{\eta}, \eta - \hat{\eta}) + O\left(\|\eta - \hat{\eta}\|^3\right). \tag{6}$$

The generalization of the first order construction now requires a von Mises type representation of the form, for measurable functions $\dot{\chi}_\eta \colon \mathcal{X} \to \mathbb{R}$ and $\ddot{\chi}_\eta \colon \mathcal{X}^2 \to \mathbb{R}$,

$$\chi'_{\hat{\eta}}(\eta - \hat{\eta}) + \tfrac{1}{2}\chi''_{\hat{\eta}}(\eta - \hat{\eta}, \eta - \hat{\eta}) = \int \dot{\chi}_{\hat{\eta}}\, d(P_\eta - P_{\hat{\eta}}) + \frac{1}{2}\iint \ddot{\chi}_{\hat{\eta}}\, d(P_\eta - P_{\hat{\eta}}) \times (P_\eta - P_{\hat{\eta}}) + O\left(\|\eta - \hat{\eta}\|^3\right). \tag{7}$$

We assume without loss of generality that the functions $\dot{\chi}_\eta$ and $\ddot{\chi}_\eta$ are degenerate relative to P_η. The von-Mises representation then suggests the "corrected plug-in estimator"

$$\chi(\hat{\eta}) + \mathbb{P}_n \dot{\chi}_{\hat{\eta}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\chi}_{\hat{\eta}}. \tag{8}$$

The empirical measure and two-sample U-statistic serve as unbiased estimators of the expectations of their kernels. For simplicity we may again base the initial estimator $\hat{\eta}$ and these two U-statistics on independent samples of observations. Because the variance of a U-statistic is of order $O(1/n)$, this estimator ought to have an error of the order $O_P(n^{-1/2}) + O_P\left(\|\eta - \hat{\eta}\|^3\right)$. We shall discuss the validity of this later.

To characterize the first and second order influence functions we can again employ smooth one-dimensional submodels $(P_{\eta_t} : t \geq 0)$. With the first and second order

derivatives of these models denoted by

$$\dot{g}_\eta = \frac{\frac{\partial}{\partial t}|_{t=0} p_{\eta_t}}{p_\eta}, \qquad \ddot{g}_\eta = \frac{\frac{\partial^2}{\partial t^2}|_{t=0} p_{\eta_t}}{p_\eta}, \tag{9}$$

the von Mises expansion (7) can informally be seen to imply

$$\frac{d}{dt}_{|t=0} \chi(\eta_t) = P_\eta \dot{\chi}_\eta \dot{g}_\eta, \tag{10}$$

$$\frac{d^2}{dt^2}_{|t=0} \chi(\eta_t) = P_\eta \dot{\chi}_\eta \ddot{g}_\eta + P_\eta^2 \ddot{\chi}_\eta (\dot{g}_\eta \times \dot{g}_\eta). \tag{11}$$

The Eq. (10) is identical to Eq. (5), and hence a first order influence function $\dot{\chi}_\eta$ can be taken as before. Following Pfanzagl (1985) we define a *second order influence function* $\ddot{\chi}_\eta$ as a measurable function $\ddot{\chi}_\eta : \mathcal{X}^2 \to \mathbb{R}$ that satisfies (11) for every path $t \mapsto \eta_t$ under consideration. From Eq. (11) we see that $\ddot{\chi}_\eta$ is unique only up to functions that are orthogonal to functions of the form $\dot{g}_\eta \times \dot{g}_\eta$, for $\dot{g}_\eta$ belonging to the tangent set. In particular, a second order influence function $\ddot{\chi}_\eta$ can always be taken to be symmetric and degenerate relative to P_η. It must be taken so in the construction of the estimator (8).

The two influence functions occur together in Eq. (11), and hence should be considered a pair $(\dot{\chi}_\eta, \ddot{\chi}_\eta)$ of functions rather than as two separate functions. This is particularly important if the tangent set is not "full", i.e., smaller than the set of all mean-zero functions in $L_2(P_\eta)$, the tangent set of a nonparametric model. Both first and second order influence functions are then non-unique, but their different versions cannot be freely combined into valid pairs $(\dot{\chi}_\eta, \ddot{\chi}_\eta)$. This is connected to the fact that first and second order derivatives $\dot{g}_\eta$ and $\ddot{g}_\eta$ are also not clearly separated. A simple change of speed $t \mapsto \phi(t)$ of a path through a second order diffeomorphism $\phi : [0, 1] \to [0, 1]$ leads to the submodel $(P_{\eta_{\phi(t)}} : t \geq 0)$ with first and second order derivatives, by the chain rule,

$$\phi'(0)\dot{g}_\eta, \qquad \phi'(0)^2 \ddot{g}_\eta + \phi''(0)\dot{g}_\eta.$$

Thus the first order derivative becomes part of the second order derivative after reparameterization. Pfanzagl (Pfanzagl 1985, 2.4.4) has shown, under assumptions of smoothness of the tangent set as a function of the parameter, that the sum of *every* first order derivative and *every* second order derivative occurs as the second order derivative of some path. Thus the set of second order derivatives $\ddot{g}_\eta$ is only defined up to equivalence modulo the tangent set.

From Eq. (11) it is also clear that second order influence functions involve the joint distribution of *two* observations. Correspondingly, we prefer to define a *second order tangent space* of the model not through the second order derivatives $\ddot{g}_\eta$ along paths

$t \mapsto p_{\eta_t}$, but through the functions of two arguments

$$\ddot{s}_\eta := \frac{\frac{\partial^2}{\partial t^2}_{|t=0}(p_{\eta_t} \times p_{\eta_t})}{p_\eta \times p_\eta} = \ddot{g}_\eta \times 1 + 2\,\dot{g}_\eta \times \dot{g}_\eta + 1 \times \ddot{g}_\eta. \tag{12}$$

The function $\ddot{s}_\eta$ is a *second order score* for the model $(P_\eta \times P_\eta : \eta \in H)$ for two observations. The corresponding first order scores are

$$\dot{s}_\eta := \frac{\frac{\partial}{\partial t}_{|t=0}(p_{\eta_t} \times p_{\eta_t})}{p_\eta \times p_\eta} = \dot{g}_\eta \times 1 + 1 \times \dot{g}_\eta. \tag{13}$$

With these notations the Eqs. (10), (11) defining the influence functions can also be written as, if $\ddot{\chi}_\eta$ is chosen degenerate,

$$\frac{d}{dt}_{|t=0} \chi(\eta_t) = P_\eta^2\left(\dot{\chi}_\eta + \tfrac{1}{2}\ddot{\chi}_\eta\right)\dot{s}_\eta = \frac{d}{dt}_{|t=0} P_{\eta_t}^2\left(\dot{\chi}_\eta + \tfrac{1}{2}\ddot{\chi}_\eta\right), \tag{14}$$

$$\frac{d^2}{dt^2}_{|t=0} \chi(\eta_t) = P_\eta^2\left(\dot{\chi}_\eta + \tfrac{1}{2}\ddot{\chi}_\eta\right)\ddot{s}_\eta = \frac{d^2}{dt^2}_{|t=0} P_{\eta_t}^2\left(\dot{\chi}_\eta + \tfrac{1}{2}\ddot{\chi}_\eta\right). \tag{15}$$

Here we interprete the function $\dot{\chi}_\eta : \mathcal{X} \to \mathbb{R}$ as a function $\dot{\chi}_\eta : \mathcal{X}^2 \to \mathbb{R}$ that depends on the first argument only (and is constant in the second), or (better) replace it by its symmetrization $(x_1, x_2) \mapsto \frac{1}{2}\left(\dot{\chi}_\eta(x_1) + \dot{\chi}_\eta(x_2)\right)$. The equations show that the overall *influence function* $\dot{\chi}_\eta + \frac{1}{2}\ddot{\chi}_\eta$ is characterized by having "correct" inner products with the overall scores $\dot{s}_\eta$ and $\ddot{s}_\eta$. This overall influence function uniquely defines its constituents $\dot{\chi}_\eta$ and $\frac{1}{2}\ddot{\chi}_\eta$ provided $\ddot{\chi}_\eta$ is restricted to be degenerate. The overall influence function is itself unique only up to projection onto the closed linear span in $L_2(P_\eta \times P_\eta)$ of all functions $\dot{s}_\eta$ and $\ddot{s}_\eta$.

The equality of the far left and right sides of Eqs. (14), (15) gives an alternative characterization of the overall influence function (at η_0) as a function such that the maps $\eta \mapsto \chi(\eta)$ and $\eta \mapsto P_\eta^2(\dot{\chi}_{\eta_0} + \frac{1}{2}\ddot{\chi}_{\eta_0})$ possess the same first and second order derivatives at η_0. Because the derivatives of a map ϕ on an open subset H of a normed space are completely characterized by the derivatives of the maps $t \mapsto \phi(\eta_0 + th)$, for h ranging over the space ("Gateaux derivatives"), we conclude that in the case of such parameters sets H it suffices to consider linear paths $t \mapsto \eta_t = \eta_0 + th$. (The mixed second derivative $\phi''_{\eta_0}(g, h)$ can be recovered from $\phi''_{\eta_0}(g + h, g + h)$ and $\phi''_{\eta_0}(g - h, g - h)$ by "polarization".) This is true more generally for parameter spaces H defined by a linear constraint, but in the case of nonlinear constraints the use of curved paths is necessary.

The plug-in estimator (8) can be written $\chi(\hat{\eta}) + \mathbb{U}_n(\dot{\chi}_{\hat{\eta}} + \frac{1}{2}\ddot{\chi}_{\hat{\eta}})$. A definition of an *efficient* or *canonical* second order influence function, should therefore refer to the variance of the U-statistic $\mathbb{U}_n(\dot{\chi}_\eta + \frac{1}{2}\ddot{\chi}_\eta)$. Unlike in the linear case this does not translate in the variance of the influence function $\dot{\chi}_\eta + \frac{1}{2}\ddot{\chi}_\eta$ itself (except for $n = 2$ if $\dot{\chi}_\eta$ is interpreted as the symmetric function $(x_1, x_2) \mapsto \frac{1}{2}\left(\dot{\chi}_\eta(x_1) + \dot{\chi}_\eta(x_2)\right)$). By

Eq. (5), if $\ddot{\chi}_\eta$ is chosen degenerate and symmetric,

$$n \operatorname{var}_\eta \mathbb{U}_n(\dot{\chi}_\eta + \tfrac{1}{2}\ddot{\chi}_\eta) = P_\eta \dot{\chi}_\eta^2 + \frac{1}{2(n-1)} P_\eta^2 \ddot{\chi}_\eta^2.$$

Thus the second order part adds a term of order $O(1/n)$ relative to the first order contribution. The norm of the function $\dot{\chi}_\eta + \frac{1}{2}\ddot{\chi}_\eta$ in $L_2(P_\eta \times P_\eta)$ is irrelevant, even though the inner product of this space determines the influence functions.

It is possible to resolve this discrepancy by working in the model with n observations. From the expansion

$$\begin{aligned}\prod_{i=1}^n \frac{p_{\eta_t}}{p_\eta}(x_i) &= \prod_{i=1}^n \left(1 + t\dot{g}_\eta(x_i) + \tfrac{1}{2}t^2 \ddot{g}_\eta(x_i) + \cdots\right) \\ &= 1 + t\sum_{i=1}^n \dot{g}_\eta(x_i) + t^2 \left(\frac{1}{2}\sum_{i=1}^n \ddot{g}_\eta(x_i) + \sum\sum_{1\le i<j\le n} \dot{g}_\eta(x_i)\dot{g}_\eta(x_j)\right) + \cdots,\end{aligned}$$

we see that first and second order scores for the model $(P_\eta^n : \eta \in H)$ take the forms

$$\dot{s}_\eta^{(n)} = \frac{\frac{\partial}{\partial t}|_{t=0}(p_{\eta_t} \times \cdots \times p_{\eta_t})}{p_\eta \times \cdots \times p_\eta} = n\mathbb{P}_n \dot{g}_\eta, \tag{16}$$

$$\ddot{s}_\eta^{(n)} = \frac{\frac{\partial^2}{\partial t^2}|_{t=0}(p_{\eta_t} \times \cdots \times p_{\eta_t})}{p_\eta \times \cdots \times p_\eta} = n\mathbb{P}_n \ddot{g}_\eta + n(n-1)\mathbb{U}_n(\dot{g}_\eta \times \dot{g}_\eta). \tag{17}$$

Rather than in the form Eqs. (14), (15), the Eqs. (10), (11) that define the influence functions can then be written in the form

$$\frac{d}{dt}_{|t=0} \chi(\eta_t) = P_\eta^n \left(\mathbb{U}_n\left(\dot{\chi}_\eta + \tfrac{1}{2}\ddot{\chi}_\eta\right)\right) \dot{s}_\eta^{(n)}, \tag{18}$$

$$\frac{d^2}{dt^2}_{|t=0} \chi(\eta_t) = P_\eta^n \left(\mathbb{U}_n\left(\dot{\chi}_\eta + \tfrac{1}{2}\ddot{\chi}_\eta\right)\right) \ddot{s}_\eta^{(n)}. \tag{19}$$

We conclude that the influence functions are determined by the inner products of the U-statistic $\mathbb{U}_n\left(\dot{\chi}_\eta + \frac{1}{2}\ddot{\chi}_\eta\right)$ in $L_2(P_\eta^n)$ with the score functions $\dot{s}_\eta^{(n)}$ and $\ddot{s}_\eta^{(n)}$. The influence functions that yield a minimal variance are found by projecting this U-statistic onto the closed linear span of these score functions. Thus it is natural to define the latter span as the *second order tangent space* of the model.

For computation in examples the defining Eq. (11) or (15) of a second order influence function can be tedious. It is usually easier to apply the rule that a second derivative is the derivative of the first derivative. In the present situation this takes the following form (Pfanzagl 1985, 4.3.11): if $\ddot{\chi}_\eta : \mathcal{X}^2 \to \mathbb{R}$ is a function such that $x_2 \mapsto \ddot{\chi}_\eta(x_1, x_2)$ is a first order influence function of the parameter $\eta \mapsto \dot{\chi}_\eta(x_1)$, for every fixed x_1 and a first order influence function $\dot{\chi}_\eta$ (not necessarily degenerate), then $\ddot{\chi}_\eta$ is a second order influence function.

Lemma 1 *Suppose that* $(P_{\eta_t}: t \geq 0)$ *is a sufficiently smooth submodel and* $\dot{\chi}_{\eta_t}: \mathcal{X} \to \mathbb{R}$ *and* $\ddot{\chi}_{\eta_t}: \mathcal{X}^2 \to \mathbb{R}$ *are measurable functions that satisfy*

$$\frac{d}{dt}\chi(\eta_t) = \int \dot{\chi}_{\eta_t} \frac{d}{dt} p_{\eta_t}\, d\mu, \qquad (t \geq 0),$$
$$\frac{d}{dt}_{|t=0} \dot{\chi}_{\eta_t}(x_1) = \int \ddot{\chi}_\eta(x_1, x_2)\dot{g}_\eta(x_2)\, dP_\eta(x_2), \qquad (x_1 \in \mathcal{X}).$$

Then the function $\ddot{\chi}_\eta$ *is a second order influence function, and so is the symmetrization of its orthogonal projection onto the degenerate functions in* $L_2(P_\eta \times P_\eta)$.

Proof By differentiation of the first identity (under the integral) we see that

$$\frac{d^2}{dt^2}\chi(\eta_t) = \int \frac{d}{dt}\dot{\chi}_{\eta_t} \frac{d}{dt} p_{\eta_t}\, d\mu + \int \dot{\chi}_{\eta_t} \frac{d^2}{dt^2} p_{\eta_t}\, d\mu.$$

We evaluate this at $t = 0$ and substitute the second identity in the first term on the right to arrive at Eq. (11). The equation remains valid if $\ddot{\chi}_\eta$ is replaced by its projection and symmetrization. □

Just as for first order influence functions there is no guarantee that a second order influence function exists. The difference is that, for the examples we are interested in, *non*existence of a second order influence function is typical. A first indication that this might happen is that the informal conclusion reached in the preceding that the quadratic estimator (8) will have an error of the order $O_P(n^{-1/2}) + O_P\left(\|\eta - \hat{\eta}\|^3\right)$ is overly optimistic. In comparison to the linear estimator (3), this estimator would have reduced the dependence on the preliminary estimator $\hat{\eta}$ from $O_P\left(\|\eta - \hat{\eta}\|^2\right)$ to $O_P\left(\|\eta - \hat{\eta}\|^3\right)$, apparently without a serious penalty on the variance of the estimator. In our examples this does not occur, simply because a second order influence function does not exist.

As for the first order influence function, the nonexistence of the second order influence function may be caused by a lack of invertibility of the map $\eta \to p_\eta$ or by failure of a von Mises type representation. The invertibility is again necessary, because we need representation of the derivatives of $\eta \mapsto \chi(\eta)$ in terms of the distribution P_η of the observation. This is similar as in the linear situation. The second cause for failure of representation also arose in the linear situation, but appears to arise in a much more serious way at the second order. Whereas a continuous, linear map $B: L_2(P_\eta) \to \mathbb{R}$ is always representable as an inner product $B(g) = P_h g \dot{\chi}_\eta$ for some function $\dot{\chi}_\eta$, a continuous, bilinear map $B: L_2(P_\eta) \times L_2(P_\eta) \to \mathbb{R}$ is not necessarily representable through a measurable function $\ddot{\chi}_\eta: \mathcal{X}^2 \to \mathbb{R}$, in the form

$$B(g, h) = \iint g(x_1)\ddot{\chi}_\eta(x_1, x_2)h(x_2)\, dP_\eta(x_1)\, dP_\eta(x_2). \tag{20}$$

It can be shown that a continuous, bilinear map can always be written in the form $B(g, h) = \int g(Ah)\, dP_\eta$ for a continuous, linear operator $A: L_2(P_\eta) \to L_2(P_\eta)$, but

the latter operator is not necessarily a *kernel operator* in that $Ah(x_1) = \int \ddot{\chi}_\eta(x_1, x_2)\, h(x_2)\, dP_\eta(x_2)$ for some *kernel* $\ddot{\chi}_\eta$. The latter representation is necessary for the von-Mises representation (7) of the second derivative.

Failure of existence of $\ddot{\chi}_\eta$ does not mean that the idea to use a quadratic expansion for improved estimation is not fruitful. Failure does mean that we cannot construct the estimator (8) and the estimation rate $O_P(n^{-1/2}) + O_P\left(\|\hat{\eta} - \eta\|^3\right)$ may not be attainable. However, we may return to Eq. (6) and try and estimate the quadratic term as well as possible, and still improve on the linear estimator. A key observation is that a bilinear map on a *finite-dimensional* subspace $L \times L \subset L_2(P_\eta) \times L_2(P_\eta)$ is always representable by a kernel.

Lemma 2 *If $L \subset L_2(P_\eta)$ is a finite-dimensional subspace and $B: L \times L \to \mathbb{R}$ is continuous and bilinear, then there exists a function $\ddot{\chi}_\eta \in L_2(P_\eta \times P_\eta)$ such that (20) holds for every $g, h \in L$.*

Proof For an arbitrary orthonormal basis $e_1, \ldots, e_k$ of L we can express an element $g \in L$ as $g = \sum_{i=1}^k \langle g, e_i\rangle_\eta e_i$, for $\langle \cdot, \cdot \rangle_\eta$ the inner product of $L_2(P_\eta)$. By bilinearity

$$\begin{aligned} B(g,h) &= \sum_{i=1}^k \sum_{j=1}^k \langle g, e_i\rangle_\eta \langle h, e_j\rangle_\eta B(e_i, e_j) \\ &= \int\int g(x_1)h(x_2) \sum_{i=1}^k \sum_{j=1}^k B(e_i, e_j) e_i(x_1) e_j(x_2)\, dP_\eta(x_1)\, dP_\eta(x_2). \end{aligned}$$

Thus the function $(x_1, x_2) \mapsto \sum_{i=1}^k \sum_{j=1}^k B(e_i, e_j) e_i(x_1) e_j(x_2)$ is a kernel for the map B. □

If the invertibility $\eta \mapsto p_\eta$ can be resolved, we can therefore always represent the second derivative in Eq. (6) at differences $\eta - \hat{\eta}$ within a given finite-dimensional linear space. The estimator (8) based on the resulting "partial second order influence function" then will add a *representation error* to the remainder $O_P\left(\|\hat{\eta} - \eta\|^3\right)$. This representation error can be made arbitrarily small by choosing the finite-dimensional linear space sufficiently large. However, the corresponding partial influence functions depend on the approximating linear spaces, the estimator now having the form

$$\chi(\hat{\eta}) + \mathbb{P}_n \dot{\chi}_{\hat{\eta}} + \tfrac{1}{2}\mathbb{U}_n \ddot{\chi}_{L,\hat{\eta}}, \tag{21}$$

where $\ddot{\chi}_{L,\eta}$ is a partial second order influence function based on an approximating space L. To obtain a good estimator we must balance the representation error, remainder $O\left(\|\hat{\eta} - \eta\|^3\right)$, and the variance of the estimator. In an asymptotic framework we let the approximating space L increase to the full space when $n \to \infty$. We shall see that this may cause the variance of $\mathbb{U}_n \ddot{\chi}_{L,\hat{\eta}}$ to dominate the variance of the linear term $\mathbb{P}_n \dot{\chi}_{\hat{\eta}}$ and the overall variance may be bigger than $O(1/n)$. However, by proper balancing of the three terms we do never worse than the linear estimator (3), and we gain over it if the parameter set H is large.

4 Estimating the square of a density

Consider the problem of estimating the functional $\chi(p) = \int p^2 \, d\mu$ based on a random sample of size n from the density p. This problem was discussed among others in Bickel and Ritov (1988) and Laurent (1996), Laurent (1997). We shall rederive the estimator by Laurent (1996) through our general approach.

As the underlying model $\mathcal{P}$ we use a set of densities that is restricted only qualitatively, for instance a Hölder space of functions on the unit square in $\mathbb{R}^d$. We parameterize this model by the density itself, which we denote by p (hence $p_\eta = \eta = p$). The tangent space of the model can then be taken equal to the set of all mean zero functions $\dot{g}_p \colon \mathcal{X} \to \mathbb{R}$ in $L_2(P)$, and the first order influence function takes the form

$$\dot{\chi}_p(x) = 2\,(p(x) - \chi(p))\,. \tag{22}$$

To see this, it suffices to note that this function is mean-zero (i.e., degenerate) and satisfies

$$\frac{d}{dt}_{|t=0} \chi(p_t) = \int 2 p_t \dot{p}_t \, d\mu_{|t=0} = P 2 p \dot{g}_p = P \dot{\chi}_p \dot{g}_p,$$

for any sufficiently regular path $t \mapsto p_t$ with $p_0 = p$ and score function $\dot{g}_p = \dot{p}_0 / p_0$ at $t = 0$. This first order influence function exists without making assumptions on p or $\mathcal{P}$.

We compute a second order influence function as the influence function of the functional $p \mapsto \bar{\chi}_p(x_1) = 2p(x_1)$, which is the first order influence function up to centering. This entails point evaluation at a fixed point x_1, which, unfortunately, is not a differentiable functional in the sense of possessing an influence function. For any sufficiently regular path $t \mapsto p_t$ with score function $\dot{g}_p$,

$$\frac{d}{dt}_{|t=0} p_t(x_1) = \dot{g}_p(x_1) p(x_1).$$

Existence of an influence function of the functional $p \mapsto p(x_1)$ would require the map $g \mapsto g(x_1)p(x_1)$ to be representable as an inner product in $L_2(P)$ on the tangent space. Such a representation is not possible (unless p has finite support), because the map is not continuous relative to the $L_2(P)$-norm.

Thus we content ourselves with partial representation of the second derivative. To this aim it is useful to think of the point evaluation map as integrating versus the Dirac measure (at x_1). Full representation of the functional $g \mapsto g(x_1)p(x_1)$ would be possible if there existed a function $\Pi \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that,

$$g(x_1)p(x_1) = \int \Pi(x_1, x_2) g(x_2) p(x_2) \, d\mu(x_2). \tag{23}$$

If this were true for every function g, then the measure $B \mapsto \int_B \Pi(x_1, x_2) \, d\mu(x_2)$ would, for each fixed x_1, act as a Dirac measure at x_1. In other words, the desired,

but not existing, function Π would be a "Dirac measure" on the diagonal of $\mathcal{X} \times \mathcal{X}$. Our second best is a function for which Eq. (23) is true, if not for all, then for a large collection of g. The kernel Π of a projection operator $\Pi\colon L_2(\mu) \to L_2(\mu)$ onto a (large) subspace is a candidate, because it satisfies the display whenever gp is in the subspace: if $\Pi f(x_1) = \int \Pi(x_1, x_2) f(x_2)\, d\mu(x_2)$, then the equation $gp = \Pi(gp)$, which is valid for every gp in the range of the projection, gives the preceding display.

Lemma 3 *An orthogonal projection $\Pi\colon L_2(\mu) \to L \subset L_2(\mu)$ onto a finite-dimensional subspace L can be represented as $\Pi f(x_1) = \int \Pi(x_1, x_2) f(x_2)\, d\mu(x_2)$ for the kernel function $\Pi(x_1, x_2) = \sum_{i=1}^k e_i(x_1) e_i(x_2)$ and $e_1, \ldots, e_k$ an orthonormal basis of L. This kernel satisfies $\int \Pi^2\, d\mu \times \mu = k$.*

Proof We have $\Pi f(x_1) = \sum_{i=1}^k \langle f, e_i\rangle_\mu e_i(x_1)$ for $\langle f, e_i\rangle_\mu = \int f e_i\, d\mu$. The representation follows by exchanging the order of summation and integration.

The square kernel is $\sum_{i=1}^k \sum_{j=1}^k e_i(x_1) e_j(x_1) e_i(x_2) e_j(x_2)$. By the orthonormality of the basis (e_i) the (double) integral of the off-diagonal terms ($i \neq j$) vanishes and the double integral of the diagonal terms is equal to 1. Thus the double integral is k. □

We also arrive at a projection operator from the formula $\chi_p''(g, h) = 2 \int gh\, p^2\, d\mu$ for the second derivative of χ. We can write this in the form $\chi_p''(g, h) = 2 \int g(A_p h)\, dP$ for the operator $A_p\colon L_2(P) \to L_2(P)$ given by $A_p h = hp$. The operator A_p is not of kernel form, but we can approximate it by ΠA_p, leading to the approximation $2 \int g(\Pi A_p h)\, dP = 2 \int gp\ (\Pi(hp))\, d\mu$ for $\chi_p''(g, h)$.

For a given orthonormal basis $e_1, e_2, \ldots$ of $L_2(\mu)$ we take the kernel $\Pi(x_1, x_2)$ of the projection onto the span of the first k elements, given by Lemma 3, as a "partial" influence function of the functional $p \mapsto p(x_1)$, and $x_2 \mapsto 2\Pi(x_1, x_2)$ as a "partial" influence function of the functional $p \mapsto \bar{\chi}_p(x_1)$. The projection of this function onto the degenerate functions is

$$\ddot{\chi}_p(x_1, x_2) = 2\Pi(x_1, x_2) - 2\Pi p(x_1) - 2\Pi p(x_2) + 2 \int (\Pi p)^2\, d\mu. \tag{24}$$

The quadratic estimator (8), given the initial estimator $\hat{p}$, takes the form

$$\begin{aligned} \chi(\hat{p}) + \mathbb{P}_n \dot{\chi}_{\hat{p}} + \mathbb{U}_n \ddot{\chi}_{\hat{p}} &= \mathbb{U}_n \Pi + \mathbb{U}_n \left((I - \Pi)\hat{p}\right) \\ &= \mathbb{U}_n \left(\sum_{i=1}^k e_i \times e_i \right) + \mathbb{U}_n \left(\sum_{i=k+1}^\infty \hat{\theta}_i e_i \right), \end{aligned}$$

for $\hat{\theta}_i = \int \hat{p} e_i\, d\mu$ the Fourier coefficients of $\hat{p}$. If we choose the initial estimator to take values in the range of Π, then $\hat{\theta}_i = 0$ for $i > k$ and the second term vanishes. The resulting estimator reduces to the estimator considered by Laurent (1996, 1997), who showed that the estimator is minimax if p is a-priori known to belong to a multiple of the unit ball in the Hölder space $C^\beta[0, 1]$ of regularity β and (e_i) is a basis suited to this a-priori model. In fact, mean and variance ofthe estimator satisfy, with θ_i the

Fourier coefficients of p,

$$\mathrm{E}_p \mathbb{U}_n \left(\sum_{i=1}^{k} e_i \times e_i \right) = \mathrm{E}_p \sum_{i=1}^{k} e_i \times e_i = \sum_{i=1}^{k} \theta_i^2,$$
$$\mathrm{var}_p \, \mathbb{U}_n \left(\sum_{i=1}^{k} e_i \times e_i \right) \leq \frac{4}{n} P(\Pi p)^2 + \frac{2k}{n(n-1)}.$$

The bound on the variance follows from Lemma 6. If it is a-priori known that $\sum_{i=1}^{\infty} \theta_i^2 i^{2\beta} < \infty$, then the bias is bounded by $\sum_{i>k} \theta_i^2 \leq k^{-2\beta}$. The square bias is balanced against the variance if k is chosen of the order $k_n = n^{2/(4\beta+1)}$ if $\beta \leq 1/4$ and $k_n = n$ if $\beta \geq 1/4$. The resulting rate of convergence is $n^{-2\beta/(4\beta+1}$ if $\beta \leq 1/4$ and $n^{-1/2}$ if $\beta \geq 1/4$. In Robins et al. (2007) it is shown that it is also asymptotically normal.

5 Estimating the mean response in missing data models

Suppose that a typical observation is distributed as $X = (YA, A, Z)$ for Y and A taking values in the two-point set $\{0, 1\}$ and conditionally independent given Z. We think of Y as a response variable, which is observed only if the indicator A takes the value 1. The covariate Z is chosen such that it contains all information on the dependence between response and missingness indicator (*missing at random*). Alternatively, we think of Y as a counterfactual outcome if a treatment were given ($A = 1$) and estimate (half) the treatment effect under the assumption of "no unmeasured confounders". Both applications may require that Z is high-dimensional (e.g., of dimension 10), and there is typically insufficient a-priori information to model the dependence of A and Y on Z.

The model can be parameterized by the marginal density f of Z (relative to some dominating measure ν) and the probabilities $b(z) = \mathrm{P}(Y = 1 | \, Z = z)$ and $a(z)^{-1} = \mathrm{P}(A = 1 | \, Z = z)$. (We use a for the inverse probability, because this simplifies later formulas.) Thus the density p_η of an observation X is described by the triple $\eta = (a, b, f)$.

We wish to estimate the mean response $\mathrm{E}Y$, i.e., the parameter

$$\chi(\eta) = \int bf \, d\nu.$$

Estimators that are $n^{-1/2}$-consistent and asymptotically efficient in the semiparametric sense have been constructed using a variety of methods (e.g., Robins and Rotnitzky 1992; van der Laan and Robins 2003; van der Vaart 1998), but only if a or b, or both, parameters are restricted to sufficiently small regularity classes. For instance, if the covariate ranges over a compact, convex subset $\mathcal{Z}$ of $\mathbb{R}^d$, then the mentioned papers provide $n^{-1/2}$-consistent estimators under the assumption that a and b belong

to Hölder classes $C^\alpha(\mathcal{Z})$ and $C^\beta(\mathcal{Z})$ with α and β large enough that

$$\frac{\alpha}{2\alpha+d}+\frac{\beta}{2\beta+d}\geq\frac{1}{2}. \tag{25}$$

For moderate to large dimensions d this is a restrictive requirement. We shall show that a quadratic estimator of the type (8) can attain a $n^{-1/2}$-rate in a bigger model and obtains a strictly better rate than the usual estimators if the $n^{-1/2}$-rate is not obtainable.

Prelimary estimators The parameter $1/a(z)=\mathrm{E}(A|\,Z=z)$ is the regression of A on Z and hence can be estimated by any nonparametric regression estimator, such as a kernel or a truncated series estimator. Similarly, the function $b(z)=\mathrm{P}(Y=1|\,Z=z,A=1)$ is the regression of the *observed* Y on Z and can be estimated by nonparametric regression based on the subsample $(Y_i\colon A_i=1)$ on the corresponding Z_i. We shall see below that the parameter f/a is more fundamental than the parameter f. By Bayes' rule $(f/a)(z)=\mathrm{P}(A=1|\,Z=z)f(z)$ is $\mathrm{P}(A=1)$ times the conditional density of Z given $A=1$. Therefore, we may estimate f/a by a nonparametric density estimator based on the subsample $(Z_i\colon A_i=1)$ times $n^{-1}\sum_{i=1}^n A_i$.

Tangent space and first order influence function The one-dimensional submodels $t\mapsto p_{\eta_t}$ induced by paths of the form $a_t=a+t\alpha$, $b_t=b+t\beta$, and $f_t=f(1+t\phi)$ for given directions α, β and ϕ yield scores $B_\eta(\alpha,\beta,\phi)=B_\eta^a\alpha+B_\eta^b\beta+B_\eta^f\phi$, for B_η^a, B_η^b, B_η^f the *score operators* for the three parameters, given by

$$\begin{aligned}
B_\eta^a\alpha(X)&=-\frac{Aa(Z)-1}{a(Z)(a-1)(Z)}\alpha(Z), && a-\text{score},\\
B_\eta^b\beta(X)&=\frac{A\,(Y-b(Z))}{b(Z)(1-b)(Z)}\beta(Z), && b-\text{score},\\
B_\eta^f\phi(X)&=\phi(Z), && f-\text{score}.
\end{aligned}$$

The first-order influence function is well known to take the form

$$\dot\chi_\eta(X)=Aa(Z)\,(Y-b(Z))+b(Z)-\chi(\eta). \tag{26}$$

To see this it must be verified that this function satisfies, for every path $t\mapsto p_{\eta_t}$ as described previously,

$$\frac{d}{dt}_{|t=0}\chi(\eta_t)=\mathrm{E}_\eta\dot\chi_\eta(X)\,B_\eta(\alpha,\beta,\phi)(X).$$

For the paths $a_t=a+t\alpha$, $b_t=b+t\beta$ and $f_t=f(1+t\phi)$ the left side of this equation is $\int(\beta+b\phi)f\,d\nu$. The right side can easily be evaluated to be the same, where it may be noted that conditional expectations of functions of Y and A given Z factorize, with $\mathrm{E}(Y-b(Z)|\,Z)=\mathrm{E}(Aa(Z)-1|\,Z)=0$ and $\mathrm{E}\left((Y-b(Z)^2|\,Z\right)=b(1-b)(Z)$.

The advantage of choosing a an inverse probability is clear from the form of the (random part of the) influence function, which is a bilinear function in (a, b). The error of the corresponding von-Mises representation can be computed to be, for a given initial estimator $\hat{\eta} = (\hat{a}, \hat{b}, \hat{f})$,

$$\chi(\hat{\eta}) - \chi(\eta) + P_\eta \dot{\chi}_{\hat{\eta}} = -\int (\hat{a} - a)(\hat{b} - b)\, \frac{f}{a}\, d\nu. \tag{27}$$

This is quadratic in the errors of the initial estimators. Actually, the form of the bias term is special in that square estimation errors of the two initial estimators $\hat{a}$ and $\hat{b}$ do no arise, but only the product of their errors. This property, termed "double robustness" in Rotnitzky and Robins (1995), Robins and Rotnitzky (2001), van der Laan and Robins (2003), makes that it suffices that one of the two parameters is estimated well. A prior assumption that the parameters a and b are α and β regular, respectively, would allow estimation errors with rates $n^{-\alpha/(2\alpha+d)}$ and $n^{-\beta/(2\beta+d)}$. If the product of these rates is $o(n^{-1/2})$, then the bias term is negligible, and the linear estimator (3) attains a rate $n^{-1/2}$. This leads to the condition (25). If this condition fails, then the "bias" (27) is greater than $O(n^{-1/2})$. The linear estimator then does not balance bias and variance and is suboptimal.

It may be noted that the marginal density f does not enter the first order influence function. Even though the functional depends on f, a rate on the initial estimator of this function is not needed for the construction of the first order estimator. This will be different at second order.

Quadratic estimator We proceed to the computation of a second order influence function using Lemma 1, by searching a function $\ddot{\chi}_\eta : \mathcal{X}^2 \to \mathbb{R}$ such that, for every $x_1 = (y_1 a_1, a_1, z_1)$, and all directions α, β, ϕ,

$$\begin{aligned} a_1\,(y_1 - b(z_1))\,\alpha(z_1) - (a_1 a(z_1) - 1)\,\beta(z_1) &= \frac{d}{dt}_{|t=0} \left[\dot{\chi}_{\eta_t}(x_1) + \chi(\eta_t)\right] \\ &= \mathrm{E}_\eta \ddot{\chi}_\eta(x_1, X_2)\, B_\eta(\alpha, \beta, \phi)(X_2). \end{aligned} \tag{28}$$

Here the expectation is relative to the variable X_2 only. Let $K_\eta : \mathcal{Z}^2 \to \mathbb{R}$ be the kernel of an operator $K_\eta : L_2(f) \to L_2(f)$ (i.e., $K_\eta g(x_1) = \int K(x_1, x_2) g(x_2) f(x_2)\, d\mu(x_2)$), and define

$$\begin{aligned} \ddot{\chi}_\eta(X_1, X_2) = &-A_1\,(Y_1 - b(Z_1))\, a(Z_2)\,(A_2 a(Z_2) - 1)\, K_\eta(Z_1, Z_2) \\ &- (A_1 a(Z_1) - 1)\, a(Z_2) A_2\,(Y_2 - b(Z_2))\, K_\eta(Z_1, Z_2). \end{aligned} \tag{29}$$

For this choice the right side of Eq. (28) can be seen to reduce to

$$a_1\,(y_1 - b(z_1))\, K_\eta \alpha(z_1) - (a_1 a(z_1) - 1)\, K_\eta \beta(z_1).$$

(Note that $\mathrm{var}\,(Aa(Z) | Z) = a(Z) - 1$.) Thus the choice Eq. (29) of $\ddot{\chi}_\eta$ satisfies Eq. (28) for every (α, β, ϕ) such that $K_\eta \alpha = \alpha$ and $K_\eta \beta = \beta$. Were K_η equal to the identity operator, then Eq. (28) would be satisfied for every (α, β, ϕ), and an exact

second order influence function would exist. Unfortunately, the identity operator is not given by a kernel. As in Sect. 4 we have to be satisfied with an influence function that gives partial representation.

To ensure that $\ddot{\chi}_\eta$ is symmetric we choose $K_\eta(z_1, z_2) = \Pi_\eta(z_1, z_2)/a(z_2)$ for Π_η a symmetric function. Specifically, we choose Π_η the kernel of an orthogonal projection $\Pi_\eta: L_2(f/a) \to L_2(f/a)$ onto a space L. The corresponding operators then (trivially) satisfy $K_\eta g = \Pi_\eta g$ for every $g \in L_2(f/a)$, and hence K_η will approximate the identity if L is large. The function (29) that results from this choice can be seen to be both symmetric and degenerate, and hence is a candidate "approximate" influence function. If S_2 symmetrizes a function of two variables (i.e., $2\,\mathrm{S}_2\, g(X_1, X_2) = g(X_1, X_2) + g(X_2, X_1)$), then this influence function can be written as

$$\ddot{\chi}_\eta(X_1, X_2) = -\,\mathrm{S}_2\left[A_1\,(Y_1 - b(Z_1))\,\Pi_\eta(Z_1, Z_2)\,(A_2 a(Z_2) - 1)\right]. \tag{30}$$

For an initial estimator $\hat{\eta}$ based on independent observations we now construct the estimator (8).

Let $\hat{\mathrm{E}}$ and $\hat{\mathrm{var}}$ denote conditional expectations given the observations used to construct $\hat{\eta}$, and let $\|\cdot\|_2$ be the norm of $L_2(f/a)$. Assume that the true functions a, f and the estimators $\hat{a}$, $\hat{f}$ are bounded away from 0 and ∞.

Theorem 1 *The estimator $\hat{\chi}_n = \chi(\hat{\eta}) + \mathbb{P}_n \dot{\chi}_{\hat{\eta}} + \frac{1}{2}\mathbb{U}_n \ddot{\chi}_{\hat{\eta}}$ with (approximate) influence functions $\dot{\chi}_\eta$ and $\ddot{\chi}_\eta$ defined by (26) and (30) with Π_η the kernel of an orthogonal projection in $L_2(f/a)$ onto a k-dimensional linear subspace satisfies*

$$\begin{aligned}\hat{\mathrm{E}}_\eta \hat{\chi}_n - \chi(\eta) &= O_P\left(\|\hat{a} - a\|_2 \|\hat{b} - b\|_2 \left\|\frac{\hat{f}}{\hat{a}} - \frac{f}{a}\right\|_2\right) \\ &\quad + O_P\left(\|a - \Pi_\eta a\|_2 \|b - \Pi_\eta b\|_2\right), \\ \hat{\mathrm{var}}_\eta \hat{\chi}_n &= O_P\left(\frac{k}{n^2} \vee \frac{1}{n}\right).\end{aligned}$$

Proof From Eqs. (27) and (30) we have

$$\begin{aligned}\hat{\mathrm{E}}\hat{\chi}_n - \chi(\eta) &= -\int (\hat{a} - a)(\hat{b} - b)\,\frac{f}{a}\,d\nu \\ &\quad -\hat{\mathrm{E}} A_1\left(Y_1 - \hat{b}(Z_1)\right)\Pi_{\hat{\eta}}(Z_1, Z_2)\left(A_2\hat{a}(Z_2) - 1\right)\Pi_{\hat{\eta}}(Z_1, Z_2) \\ &= -\int (\hat{a} - a)(\hat{b} - b)\,\frac{f}{a}\,d\nu + \iint \left((\hat{a} - a) \times (\hat{b} - b)\right) \\ &\quad \times \Pi_{\hat{\eta}}\left(\frac{f}{a} \times \frac{f}{a}\right)\,d\nu \times \nu.\end{aligned}$$

The double integral on the far right with $\Pi_{\hat{\eta}}$ replaced by Π_η can be written as the single integral $\int (\hat{a} - a)\Pi_\eta(\hat{b} - b)\,(f/a)\,d\nu$. Added to the first integral on the right

this gives

$$-\int (\hat{a}-a)(I-\Pi_\eta)(\hat{b}-b)\,(f/a)\,dv.$$

By the Cuachy-Schwarz inequality this is bounded in absolute value by the second term in the upper bound for the bias.

Replacement of $\Pi_{\hat{\eta}}$ by Π_η in the double integral gives a difference

$$\begin{aligned}&\iint \Big((\hat{a}-a)\times(\hat{b}-b)\Big)\,(\Pi_{\hat{\eta}}-\Pi_\eta)\left(\frac{f}{a}\times\frac{f}{a}\right)\,dv\times v\\&=\int(\hat{a}-a)\left(\Pi_{\hat{\eta}}\left((\hat{b}-b)\frac{f/a}{\hat{f}/\hat{a}}\right)-\Pi_\eta(\hat{b}-b)\right)\frac{f}{a}\,dv.\end{aligned}$$

By the Cauchy-Schwarz inequality the absolute value of this is bounded above by

$$\|\hat{a}-a\|_2\left\|(\Pi_{\hat{\eta}}\circ M_{\hat{w}}-\Pi_\eta)(\hat{b}-b)\right\|_{2,\hat{v}}\|\hat{w}\|_\infty.$$

Here $M_{\hat{w}}$ is multiplication by the function $\hat{w}=(f/a)/(\hat{f}/\hat{a})$ (defined by $M_{\hat{w}}g=g\hat{w}$), and $\|\cdot\|_{2,\hat{v}}$ is the $L_2(\hat{v})$-norm for the measure $\hat{v}$ defined by $d\hat{v}=(\hat{f}/\hat{a})\,dv$. Considering $\Pi_{\hat{\eta}}$ as the projection in $L_2(\hat{v})$ with weight 1, and Π_η as the weighted projection in $L_2(\hat{v})$ with weight function $\hat{w}$, we can apply Lemma 4 to the middle term and conclude that this is bounded in absolute value by $\|\Pi_\eta\|_{2,\hat{v}}\|\hat{w}-1\|_\infty\|\hat{b}-b\|_{2,\hat{v}}$. Because we assume that the functions f/a and $\hat{f}/\hat{a}$ are bounded away from zero and infinity, this can be seen to yield the first term in the upper bound on the bias.

The function $\dot{\chi}_{\hat{\eta}}$ is uniformly bounded and hence the (conditional) variance of $\mathbb{P}_n\dot{\chi}_{\hat{\eta}}$ is of the order $O_P(1/n)$. Thus for the variance bound it suffices to consider the (conditional) variance of $\mathbb{U}_n\ddot{\chi}_{\hat{\eta}}$. In view of Lemma 6 this is bounded above by

$$\tfrac{4}{n}\mathrm{E}_\eta\mathrm{E}_\eta\left(\ddot{\chi}_{\hat{\eta}}(X_1,X_2)|\,X_1\right)^2+\tfrac{2}{n(n-1)}\mathrm{E}_\eta\ddot{\chi}^2_{\hat{\eta}}(X_1,X_2).$$

The variables $A\left(Y-\hat{b}(Z)\right)$ and $\left(A\hat{a}(Z)-1\right)$ are uniformly bounded. Hence the last term on the right is bounded above by a multiple of $n^{-2}\int\int\Pi^2_{\hat{\eta}}\,(f/a\times f/a)\,dv\times v$, which is bounded by $\|\hat{w}\|^2_\infty k/n^2$, by Lemma 3. The first order term is of the order $O(1/n)$. To see this we first note that

$$\begin{aligned}\mathrm{E}_\eta\left(\ddot{\chi}_{\hat{\eta}}(X_1,X_2)|\,X_1\right)&=-A_1\left(Y_1-\hat{b}(Z_1)\right)\Pi_{\hat{\eta}}\left((\hat{a}-a)\hat{w}\right)(X_1)\\&\quad-\left(A_1\hat{a}(Z_1)-1\right)\Pi_{\hat{\eta}}\left((\hat{b}-b)\hat{w}\right)(X_1).\end{aligned}$$

Here the variables $A_1\left(Y_1-\hat{b}(Z_1)\right)$ and $\left(A_1\hat{a}(Z_1)-1\right)$ are uniformly bounded, and the second moment of $\Pi_{\hat{\eta}}g$ is bounded by $\|\hat{w}\|_\infty$ times the second moment of g in $L_2(\hat{v})$, for every g. □

Conclusion Assume that the parameters a, b and f/a are known to be "regular" of degrees α, β and ϕ, respectively, in the sense that there exists a sequence of k-dimensional linear spaces L_k such that, for some constant C,

$$\|a - L_k\|_2 \le C\left(\frac{1}{k}\right)^{\alpha/d}, \quad \|b - L_k\|_2 \le C\left(\frac{1}{k}\right)^{\beta/d}, \quad \left\|\frac{f}{a} - L_k\right\|_2 \le C\left(\frac{1}{k}\right)^{\beta/d}.$$

This is true, for instance, if the functions a, b and f/a are defined on a compact, convex domain in $\mathbb{R}^d$ and are known to belong to Hölder (or Besov) spaces of functions of smoothness α, β and ϕ. The approximation is then valid even with the uniform norm on the left side, where the spaces L_k can be taken to be generated by polynomials, splines or wavelets.

In this case there also exist estimators $\hat{a}$ and $\hat{b}$ and $\hat{f}/\hat{a}$ that achieve convergence rates $n^{-\alpha/(2\alpha+d)}$, $n^{-\beta/(2\beta+d)}$ and $n^{-\phi/(2\phi+d)}$, respectively, uniformly over these a-priori models. Then the estimator $\hat{\chi}_n$ of Theorem 1 attains the square rate of convergence

$$\frac{k}{n^2} \vee \frac{1}{n} \vee \left(\frac{1}{n}\right)^{2\alpha/(2\alpha+d)+2\beta/(2\beta+d)+2\phi/(2\phi+d)} \vee \left(\frac{1}{k}\right)^{(2\alpha+2\beta)/d}.$$

The optimal value of k balances the first and fourth terms and is of the order $k \sim n^{2d/(d+2\alpha+2\beta)}$. The resulting rate is $n^{-\gamma}$ for

$$\gamma = \left(\frac{1}{2}\right) \wedge \left(\frac{\alpha}{2\alpha+d} + \frac{\beta}{2\beta+d} + \frac{\phi}{2\phi+d}\right) \wedge \left(\frac{2\alpha+2\beta}{d+2\alpha+2\beta}\right).$$

This reduces to the rate $n^{-1/2}$ under condition (25), but also if $(\alpha+\beta)/2 \ge d/4$ and ϕ is sufficiently large:

$$\frac{\phi}{2\phi+d} \ge \frac{1}{2} - \frac{\alpha}{2\alpha+d} - \frac{\beta}{2\beta+d}.$$

(In this case we can also choose $k = n$ independent of α and β.) In case the rate $n^{-\gamma}$ is slower than $n^{-1/2}$, then it is still better than the rate $n^{-\alpha/(2\alpha+d)-\beta/(2\beta+d)}$ obtained by the linear estimator (3).

Thus the quadratic estimator outperforms the linear estimator.

6 Technical results

Let L be a given closed subspace of $L_2(\mathcal{X}, \mathcal{A}, \mu)$ and $w: \mathcal{X} \to \mathbb{R}$ a bounded, measurable function. Define operators $\Pi, \Pi_w: L_2(\mu) \to L_2(\mu)$ by

$$\Pi g = \operatorname*{argmin}_{l \in L} \int (g-l)^2 \, d\mu,$$

$$\Pi_w g = \operatorname*{argmin}_{l \in L} \int (g-l)^2 \, w \, d\mu.$$

Thus Π is the ordinary orthogonal projection on the space L, and Π_w is a *weighted* projection. The projections can be characterized by the orthogonality relationships $\int (g - \Pi g) l \, d\mu = 0$ and $\int (g - \Pi g) l \, w \, d\mu = 0$, for every $l \in L$.

Lemma 4 *Let Π_w and Π be the weighted projections onto a fixed subspace L of $L_2(\mu)$ relative to the weight functions w and 1, respectively, and let M_w be multiplication by the function w. Then $\|\Pi_w - \Pi M_w\|_2 \leq \|\Pi_w\|_2 \|w - 1\|_\infty$.*

Proof The orthogonality relationships for the projections Π and Π_w imply that, for every $l \in L$ and g,

$$\int \Pi(wg) l \, d\mu = \int wgl \, d\mu = \int w(\Pi_w g) l \, d\mu.$$

Because $\Pi_w g - \Pi(wg)$ is contained in L,

$$\begin{aligned} \|\Pi_w g - \Pi(wg)\|_2^2 &= \int (\Pi_w g - \Pi(wg)) \, (\Pi_w g - \Pi(wg)) \, d\mu, \\ &= \int (\Pi_w g - \Pi(wg)) \, (\Pi_w g - (\Pi_w g) w) \, d\mu. \end{aligned}$$

An application of the Cauchy-Schwarz inequality and next cancellation of one factor $\|\Pi_w g - \Pi(wg)\|_2$ gives that $\|\Pi_w g - \Pi(wg)\|_2 \leq \|(\Pi_w g)(1 - w)\|_2$. The right side is bounded above by $\|\Pi_w\|_2 \|g\|_2 \|1 - w\|_\infty$. □

Lemma 5 *For degenerate, symmetric functions $f, g: \mathcal{X}^2 \to \mathbb{R}$ we have $P^n \mathbb{U}_n f = 0$ and*

$$P^n (\mathbb{U}_n f)(\mathbb{U}_n g) = \frac{1}{\binom{n}{2}} P^2 fg.$$

Lemma 6 *For any measurable function $f: \mathcal{X}^2 \to \mathbb{R}$, and $f_1(x_1) = \int f(x_1, x_2) \, dP(x_2)$,*

$$\operatorname{var} \mathbb{U}_n f \leq \frac{4}{n} P f_1^2 + \frac{2}{n(n-1)} P f^2.$$

Proof The first lemma follows by writing the square sum $(\mathbb{U}_n f)^2$ as a double sum (over ordered pairs $i < j$). The expected values of the off-diagonal terms vanish by degeneracy.

For a general measurable function $f: \mathcal{X}^2 \to \mathbb{R}$ the mean Pf^2 is the projection onto the constant functions, and the function $\bar{f}_1$ defined by $\bar{f}_1(x_1) = \int f(x_1, x_2) \, dP(x_2) - P^2 f$ is the projection of f in $L_2(P^2)$ onto the mean zero functions of one variable. The decomposition

$$f(x_1, x_2) = P^2 f + \bar{f}_1(x_1) + \bar{f}_1(x_2) + f_{12}(x_1, x_2),$$

where f_{12} is defined by the equation yields the Hoeffding decomposition $\mathbb{U}_n f = P^2 f + 2\mathbb{P}_n \bar{f}_1 + \mathbb{U}_n f_{12}$ of the U-statistic in orthogonal parts, with $\mathbb{U}_n f_{12}$ degenerate. Using Lemma 5 we see that the variance of $\mathbb{U}_n f$ is equal to $(4/n)P\bar{f}_1^2 + 2/(n(n-1))P^2 f_{12}^2$. The norm of $\bar{f}_1$ is smaller than the norm of f_1. Because f_{12} is a projection of f, its norm is bounded by the norm of f. □

References

Begun JM, Hall WJ, Huang WM, Wellner JA (1983) Information and asymptotic efficiency in parametric–nonparametric models. Ann Stat 11(2):432–452

Bickel PJ, Ritov Y (1988) Estimating integrated squared density derivatives: sharp best order of convergence estimates. Sankhyā Ser A 50(3):381–393

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation for semiparametric models. Johns Hopkins series in the mathematical sciences. Johns Hopkins University Press, Baltimore

Bolthausen E, Perkins E, van der Vaart A (2002) Lectures on probability theory and statistics. In: Lecture notes in mathematics, Lectures from the 29th summer school on probability theory held in Saint-Flour, vol 1781. Springer, Berlin. July 8–24, 1999 (edited by Bernard P)

Emery M, Nemirovski A, Voiculescu D (2000) Lectures on probability theory and statistics, In: Lecture notes in mathematics, Lectures from the 28th summer school on probability theory held in Saint-Flour, vol. 1738. Springer, Berlin. 17 August–3 September, 1998 (edited by Bernard P)

Klaassen CAJ (1987) Consistent estimation of the influence function of locally asymptotically linear estimators. Ann Stat 15(4):1548–1562

Koševnik JA, Levit BJ (1976) On a nonparametric analogue of the information matrix. Teor Verojatnost i Primenen 21(4):759–774

Laurent B (1996) Efficient estimation of integral functionals of a density. Ann Stat 24(2):659–681

Laurent B (1997) Estimation of integral functionals of a density and its derivatives. Bernoulli 3(2):181–211

Le Cam L (1960) Locally asymptotically normal families of distributions. Certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses. Univ california Publ Stat 3:37–98

Murphy SA, van der Vaart AW (2000) On profile likelihood. J Am Stat Assoc 95(450):449–485 (with comments and a rejoinder by the authors)

Pfanzagl J (1982) Contributions to a general asymptotic statistical theory. In: Lecture notes in statistics, vol 13. Springer, New York (with the assistance of Wefelmeyer W)

Pfanzagl J (1985) Asymptotic expansions for general statistical models. In: Lecture notes in statistics, vol 31. Springer, Berlin (with the assistance of Wefelmeyer W)

Robins J, Rotnitzky A (2001) Comment on the bickel and kwon article, "inference for semiparametric models: Some questions and an answer". Stat Sin 11(4):920–936

Robins J, Li L, Tchetgen E, van der Vaart A (2007) Asymptotic normality of quadratic estimators. Ann Stat (submitted)

Robins JM, Rotnitzky A (1992) Recovery of information and adjustment for dependent censoring using surrogate markers pp 297–331

Rotnitzky A, Robins JM (1995) Semi-parametric estimation of models for means and covariances in the presence of missing data. Scand J Stat 22(3):323–333

van der Laan MJ, Robins JM (2003) Unified methods for censored longitudinal data and causality. Springer Series in Statistics. Springer, New York

van der Vaart A (1991) On differentiable functionals. Ann Stat 19(1):178–204

van der Vaart A (1994) Maximum likelihood estimation with partially censored data. Ann Stat 22(4):1896–1916

van der Vaart AW (1988) Statistical estimation in large parameter spaces, CWI Tract, vol 44. Stichting Mathematisch Centrum Centrum voor Wiskunde en Informatica, Amsterdam

van der Vaart AW (1998) Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics., vol 3 Cambridge University Press, Cambridge

Von Mises R (1947) On the asymptotic distribution of differentiable statistical functions. Ann Math Stat 18:309–348

Wellner JA, Klaassen CAJ, Ritov Y (1993) Semiparametric models: a review of progress since BKRW. In: Frontiers in statistics, pp 25–44. Imp Coll Press, London (2006)